

Resistant Multiple Sparse Canonical Correlation

Jacob Coleman, Joseph Replogle, Gabriel Chandler, and Johanna Hardin

December 22, 2015

Abstract

Canonical Correlation Analysis (CCA) is a multivariate technique that takes two datasets and forms the most highly correlated possible pairs of linear combinations between them. Each subsequent pair of linear combinations is orthogonal to the preceding pair, meaning that new information is gleaned from each pair. By looking at the magnitude of coefficient values, we can find out which variables can be grouped together, thus better understanding multiple interactions that are otherwise difficult to compute or grasp intuitively.

CCA appears to have quite powerful applications to high throughput data, as we can use it to discover, for example, relationships between gene expression and gene copy number variation. One of the biggest problems of CCA is that the number of variables (often upwards of 10,000) makes biological interpretation of linear combinations nearly impossible. To limit variable output, we have employed a method known as Sparse Canonical Correlation Analysis (SCCA), while adding estimation which is resistant to extreme observations or other types of deviant data. In this paper, we have demonstrated the success of resistant estimation in variable selection using SCCA. Additionally, we have used SCCA to find *multiple* canonical pairs for extended knowledge about the datasets at hand. Again, using resistant estimators provided more accurate estimates than standard estimators in the multiple canonical correlation setting.

1 Introduction

High-throughput data is infamous for having myriad complicated functional relationships both within and across different types of measurements on the same samples. Multivariate statistics have been useful to understand molecular relationships by applying and modifying such techniques as principal component analysis (Pearson, 1901; Zou et al., 2006) and partial least squares (Nguyen and Rocke, 2001; Wold, 1973). Canonical correlation is another multivariate statistical technique used to relate two datasets evaluated on the same samples. Recent work includes Wang et al. (2014) and Hong et al. (2013), who use sparse canonical correlation to infer gene networks as tightly connected groups. Lê Cao et al. (2009) use sparse canonical correlation to compare two different microarray platforms. Gao et al. (2014) provides theoretical justification of the use of sparse canonical correlation in practice.

Canonical Correlation Analysis (CCA), proposed by Hotelling (1936), is a multivariate method for finding linear combinations of variables in high dimensions. Given two data sets, (traditional) CCA produces as many pairs of linear combinations - called canonical pairs - as variables in the smaller set. Each canonical pair has an associated correlation, called canonical correlation, and is orthogonal to every other pair. The canonical pairs, derived through singular value decomposition of the joint covariance matrix, are ordered by their associated canonical correlations. The goal of CCA is to maximize the canonical correlations.

While CCA is extremely useful for efficiently discerning relationships between variables, there are some drawbacks. Sensitivity to noise and outliers is one problem of CCA, and resistant CCA has only had minimal exploration in the literature (Branco et al., 2005; Karmel, 1991). Especially in high dimensions, even a small amount of noise or outlying values can lead to falsely high correlations and incorrectly associated variables. To address this, we use Spearman correlations to create both correlation and covariance measures. Through simulation, we demonstrate the need for and success of using a Spearman-like covariance estimate during CCA. While resistant CCA might find pairs of the most highly correlated linear combinations, variable selection is somewhat limited because the output includes coefficients for *every* variable in both datasets.

Particularly if the number of variables is quite large, if the goal is to find highly correlated groups of variables, CCA becomes impractical. That is, a linear combination of thousands of variables is difficult to interpret, and the analyst will be unable to discern which variables are most important. To handle the large number of coefficients reported from CCA, we employ a technique known as Sparse Canonical Correlation Analysis (SCCA) which sets some of the coefficients to zero. Parkhomenko et al. (2009) introduce SCCA and provide an algorithm for computing sparse variables, and subsequently demonstrate the success of SCCA for variable selection with a latent variable simulation model. Parkhomenko et al. (2009) also demonstrate that as sample size decreases, SCCA outperforms CCA. Using a similar technique but from the perspective of Penalized Matrix Decomposition, Witten et al. (2009) also explore SCCA and provide the framework for computing sparse variables with different penalty functions. In investigations of

extensions of SCCA, Chalise and Fridley (2011) explore different penalty functions and their relative successes on simulated data. We use an algorithm similar to Parkhomenko et al. (2009) with two modifications to first create a resistant measure and then subsequently to extend the method to find multiple canonical pairs.

As with (traditional) CCA, when using SCCA to analyze two different types of data (e.g., phenotypic and genotypic), there is typically interest in not only the first canonical relationship, but also in secondary relationships. Using related techniques to Principal Component Analysis (PCA) where the observations are transformed into multiple linear relationships, CCA also partitions the data into linear subspaces where multiple pairs of linear relationships describe the existing patterns. We use singular value decomposition on the cross covariance matrix to find sequential canonical pairs which are highly correlated. Note that Witten and Tibshirani (2009) briefly mention one idea for extending SCCA to MSCCA, but they do not assess the method or give the reader a sense of how to find the number of canonical pairs which should be considered significant. None of the other references using or extending SCCA consider the case of more than one canonical pair.

In section 2, we present the background mathematics of CCA (Hotelling, 1936), SCCA (Chalise and Fridley, 2011; Parkhomenko et al., 2009; Witten et al., 2009), and our derivation of multiple SCCA (MSCCA) and resistant multiple SCCA (RMSCCA). We then present our results in a series of simulations. In subsection 2.5, we describe our process for establishing a cutoff for determining which canonical pairs are significant. Our results comparing MSCCA and RM-

SCCA are given in subsection 3.2. RMSCCA is applied to publicly available data in subsection 3.3. We conclude our work in section 4.

2 Mathematical Derivation of RMSCCA

2.1 Derivation of Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) derives pairs of linear combinations between two distinct datasets that are as highly correlated as possible. The focus of CCA is to reveal relationships both within one group of variables and between the two sets of variables; coefficient values in one linear combination explain relationships within one dataset, while the pair of linear combinations explains relationships between datasets. Though canonical correlation does not distinguish between the explanatory and response variables, CCA can be considered in the context of multivariate regression. First developed by Hotelling (1936), CCA is a powerful tool for quickly determining relationships between a large number of variables. The output of CCA will be pairs of linear combinations ordered by correlation between linear combinations, such that each linear combination is orthogonal to every preceding linear combination. The coefficients for the linear combinations are called *canonical vectors*, while the linear combinations themselves are called *canonical variables*. The correlations between the *canonical variables* are called *canonical correlations*.

Consider a pair of datasets, $\mathbf{x}_{n \times p}$ and $\mathbf{y}_{n \times q}$ where the columns are variables and the rows are observations, CCA finds linear combinations of the p -dimensional random vector \mathbf{X} and the q -dimensional random vector \mathbf{Y} . The canonical vectors

α and β maximize

$$\rho(\alpha'X, \beta'Y) = \frac{\alpha'\Sigma_{XY}\beta}{\sqrt{\alpha'\Sigma_{XX}\alpha\beta'\Sigma_{YY}\beta}},$$

where

$$Cov(X, Y) = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \quad (1)$$

Because the scaling of α and β does not affect the maximum correlation, CCA returns canonical vectors subject to the additional constraint:

$$\max_{\alpha, \beta} \frac{\alpha'\Sigma_{XY}\beta}{\sqrt{\alpha'\Sigma_{XX}\alpha\beta'\Sigma_{YY}\beta}} \text{ subject to } \alpha'\Sigma_{XX}\alpha = \beta'\Sigma_{YY}\beta = 1. \quad (2)$$

Once the first linear combinations are found (called the *first canonical pair*), CCA maximizes the correlation between pairs of linear combinations of X and Y under the constraint that the second pair of linear combinations is orthogonal to the first. The process is repeated $\min(p, q)$ times.

The canonical correlation algorithm can be reduced to a singular value decomposition problem where α and β are the right and left singular vectors of

$$K = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2} = UDV^T \quad (3)$$

with $U = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k)$ and $V = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$ and k is the rank of the matrix K . When using real data, both Σ_{XX} and Σ_{YY} are estimated using the diagonal of the sample covariance matrices, as done by Chalise and Fridley (2011);

Parkhomenko et al. (2009); Witten and Tibshirani (2009). The i^{th} canonical pair can then be represented by the i^{th} singular vectors, where the canonical vectors are given by

$$\boldsymbol{\alpha}_i = \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{X}}^{-1/2} \mathbf{u}_i, \quad (4)$$

$$\boldsymbol{\beta}_i = \boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}_i. \quad (5)$$

2.2 Derivation of SCCA

As with CCA, Sparse Canonical Correlation (SCCA) is a method used to create canonical vectors which represent linear combinations of two distinct datasets. Additionally, SCCA is also based on singular value decomposition (SVD) of the covariance matrix. Because SVD can be thought of as an iterative algorithm to find the singular vectors which lead to the decomposition, Parkhomenko et al. (2009) use a SVD-like algorithm with an additional thresholding parameter to control the number of variables included in the solution of the canonical vector. The thresholding is a form of L_1 -regularization similar to LASSO (Tibshirani, 1996) which sets small values of the coefficients to zero. The algorithm due to Parkhomenko et al. (2009) is for the *first canonical pair* and is given as follows.

Algorithm 1. *Let λ_u and λ_v be chosen. Select initial values of \mathbf{u}^0 and \mathbf{v}^0 . Set $i = 0$ and K to be as given in equation (3) (i indexes the canonical pairs).*

1. *Update \mathbf{u} :*

$$(a) \quad \mathbf{u}^{i+1} \leftarrow K \mathbf{v}^i$$

$$(b) \quad \text{Normalize: } \mathbf{u}^{i+1} \leftarrow \frac{\mathbf{u}^{i+1}}{\|\mathbf{u}^{i+1}\|}$$

(c) *Soft-thresholding for sparse solution:*

$$\mathbf{u}_j^{i+1} \leftarrow (|\mathbf{u}_j^{i+1}| - \frac{1}{2}\lambda_u)_+ \text{Sign}(\mathbf{u}_j^{i+1}) \text{ for } j = 1, \dots, p$$

(d) *Normalize:* $\mathbf{u}^{i+1} \leftarrow \frac{\mathbf{u}^{i+1}}{\|\mathbf{u}^{i+1}\|}$

2. *Update \mathbf{v} :*

(a) $\mathbf{v}^{i+1} \leftarrow K\mathbf{u}^{i+1}$

(b) *Normalize:* $\mathbf{v}^{i+1} \leftarrow \frac{\mathbf{v}^{i+1}}{\|\mathbf{v}^{i+1}\|}$

(c) *Soft-thresholding for sparse solution:*

$$\mathbf{v}_j^{i+1} \leftarrow (|\mathbf{v}_j^{i+1}| - \frac{1}{2}\lambda_v)_+ \text{Sign}(\mathbf{v}_j^{i+1}) \text{ for } j = 1, \dots, p$$

(d) *Normalize:* $\mathbf{v}^{i+1} \leftarrow \frac{\mathbf{v}^{i+1}}{\|\mathbf{v}^{i+1}\|}$

3. $i \leftarrow i + 1$

4. *Repeat steps 1-3 until convergence.*

where $(x)_+$ is equal to x if $x \geq 0$ and 0 if $x < 0$ and

$$\text{Sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0. \end{cases}$$

Also, define the norm of a vector \mathbf{y} as $\|\mathbf{y}\| = \sqrt{\mathbf{y}^T \mathbf{y}}$.

We follow the convention of Parkhomenko et al. (2009) to both set the initial canonical coefficient vectors (\mathbf{u}^0 and \mathbf{v}^0) to be the row means and column means, respectively, of the K matrix and to use cross-validation to find optimal values of λ_u and λ_v . Because our work concerns finding multiple canonical pairs, our cross validation scheme is derived in the next section on MSCCA. Note that \mathbf{u}

and \mathbf{v} are sparse, and in an actual data analysis, we use diagonal versions of Σ_{XX} and Σ_{YY} . Therefore, the canonical vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, represented by equations (4) and (5), are also sparse.

2.3 Derivation of MSCCA

In CCA, sequential canonical coefficient vectors are found to simultaneously maximize the relevant correlation while maintaining orthogonality with the previous canonical coefficient vectors. With sparse vectors, when maximizing the canonical correlation, it is necessary to choose between orthogonality and sparsity. In order to address the sparse / orthogonality tradeoff, the singular value decomposition can be adapted to accommodate the information reduction after the first canonical pair is found. Recall that CCA is based on SVD of the scaled cross-covariance matrix, as in equation (3). It can be shown that the matrix K can be further decomposed into singular vectors and variables.

$$\begin{aligned} K &= UDV^T \\ &= d_1 \mathbf{u}_1 \mathbf{v}_1^T + d_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots d_k \mathbf{u}_k \mathbf{v}_k^T \end{aligned}$$

where d_i is the i^{th} singular value. Note that because the canonical vectors are orthogonal, the first singular value can be written as a function of the singular vectors and the matrix K .

$$\mathbf{u}_1^T K \mathbf{v}_1 = d_1.$$

Using the ideas above for SVD, we extend the result to get the following recursive relationship:

$$K_{i+1} = K_i - (\mathbf{u}_i^T K_i \mathbf{v}_i) \mathbf{u}_i \mathbf{v}_i^T$$

Each computation of the i^{th} canonical pair will be based on using K_i in Algorithm 1.

Witten and Tibshirani (2009) mention extending SCCA to MSCCA, and use a similar derivation to the one we have provided above. However, their SCCA algorithm is slightly different, and they provide no guidance for how to choose the number of significant pairs of canonical relationships.

Important to Algorithm 1 is the choice of λ_u and λ_v . The thresholding values should be optimal for a given dataset but should not overfit the data. Note that the values of λ_u and λ_v for the first canonical pair will impact the decomposition of K for the next canonical pair (and for all following canonical pairs). The goal of Algorithm 2 is to find the optimal values of λ_u and λ_v for each canonical pair.

Algorithm 2. *Let $\mathbf{x}_{n \times p}$ and $\mathbf{y}_{n \times q}$ represent two distinct datasets. Set $i = 1$. Split the data into $n.cv$ cross validation partitions. This creates $n.cv$ test sets, where the training set consists of all data not included in the particular partition. Let $K_1 = \Sigma_{\mathbf{X}\mathbf{X}}^{-1/2} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2}$. Again, when using real data, both $\Sigma_{\mathbf{X}\mathbf{X}}$ and $\Sigma_{\mathbf{Y}\mathbf{Y}}$ are estimated using the diagonal of the sample covariance matrices, as done by Chalise and Fridley (2011); Parkhomenko et al. (2009); Witten and Tibshirani (2009).*

1. *Compute λ_u and λ_v for i^{th} canonical pair:*

- (a) Let λ_u and λ_v range separately along a grid of points in an interval $I_\lambda \in [0, 2]$.
- (b) For each (λ_u, λ_v) pair, use Algorithm 1 to find the canonical vectors and related canonical correlations (denoted cc_{test}) on the test data.
- (c) Repeat step 1 (b) for each of the $n.cv$ choices for the test data.
- (d) Choose as $(\lambda_u^*, \lambda_v^*)$ the pair of thresholding variables that maximize the average canonical correlation (\overline{cc}_{test}) of the training canonical vectors applied to the test data (averaged over the $n.cv$ test data partitions).
- (e) Using $(\lambda_u^*, \lambda_v^*)$ and Algorithm 1, find the canonical vectors based on the entire dataset to find \mathbf{u}^* and \mathbf{v}^* .

2. Adjust K :

$$K_{i+1} = K_i - (\mathbf{u}_i^{*T} K_i \mathbf{v}_i^*) \mathbf{u}_i^* \mathbf{v}_i^{*T}$$

3. $i \leftarrow i + 1$

4. Let pq^* be the number of desired canonical pairs. Repeat Steps 1-3 for $pq^* \leq \min(p, q)$ canonical pair values.

5. Output consists of

- (a) A $pq^* \times 2$ matrix of $(\lambda_u^*, \lambda_v^*)$ pairs which have maximized correlations based on training data.
- (b) A pair of canonical correlations for each of the pq^* canonical pairs: cc on the full data set, and \overline{cc}_{test} , the average over all test sets for $(\lambda_u^*, \lambda_v^*)$.

2.4 Resistant MSCCA

In previous work on SCCA, estimation of the covariance matrix (see Equation (1)) has been done using maximum likelihood estimation. Using maximum likelihood estimation is akin to maximizing the Pearson correlation in finding thresholding variables and canonical variables. There has been some work in the literature on robust CCA, but, for example, Branco et al. (2005) considers only situations with p and q as large as 4; Dehon et al. (2000) considers only p and q as large as 3.

Because biological high-throughput data (and other types of data in high dimensions to which canonical correlation and its variants are often applied) are notoriously noisy, we give results on a resistant version of MSCCA applied to both multivariate normal data as well as heavy tailed data. The methods and algorithms above are as given in the preceding algorithms except that the covariance matrices (Equation (1)) are calculated based on the ranked data as given in the `cov(., ., method="spearman")` function in R (R Core Team, 2014). Due to the computational complexity of the algorithms, we have used a simple resistance measure. If the user has an ability to parallelize the complete application of the algorithm, it would be worth considering other estimates of covariance like the minimum covariance determinant (Rousseeuw, 1984), projection pursuit (Huber, 1985), or M-estimates (Hardin et al., 2007).

2.5 Simulating Significance Cutoff

An important step in using multiple canonical correlation pairs is deciding the number of canonical pairs to consider as significant. In order to address concerns

about multiple comparisons, we use a permutation scheme (100 permutations in the simulations below) that provides a correlation cutoff which controls the overall level of significance. Because the process for finding canonical coefficients optimizes the respective correlation, the first canonical correlation value tends to be quite high. Similarly, the subsequent correlations are typically decreasing but are often higher than standard correlations on most datasets. Therefore, it is important to have a method which evaluates which canonical pairs are significant while controlling for familywise error rate.

For analyzing an actual (or simulated) dataset, we wrote the following algorithm.

Algorithm 3. *Let $\mathbf{x}_{n \times p}$ and $\mathbf{y}_{n \times q}$ represent two distinct datasets. Set $i = 1$, (i indexes the canonical pair). Let $n.perm$ be the number of permutations.*

1. *Canonical correlation values on permuted data:*
 - (a) *Permute the rows of (WLOG) \mathbf{y} .*
 - (b) *Apply Algorithms 1 and 2 to the permuted data to find the $pq^* \leq \min(p, q)$ canonical correlations.*
 - (c) *Repeat steps 1(a) and 1(b) for $n.perm$ permutations of the original data.*
 - (d) *Let $\overline{cc}_{perm,i,(Q)}$ be the Q^{th} percentile (averaged test data) correlation (across $n.perm$ correlations) for the i^{th} canonical pair.*
2. *Apply Algorithms 1 and 2 to the original data to find the pq^* canonical correlations, cc_i for the i^{th} canonical pair.*
3. *Finding significant correlations:*

(a) Let j^* be the largest value of j such that:

$$\overline{cc}_{j,test} > \overline{cc}_{perm,j,(Q)} \quad \forall j \leq j^*$$

4. Report the canonical variables and respective canonical correlations on the original data from 1 to j^* .

The algorithm allows us to report the top j^* canonical pairs as significant. Because all j^* correlations are above the pointwise Q quantile of the permutation scheme, we have controlled our familywise error rate at $100 - (Q)\%$ (see subsection 3.2.4).

The reason that the comparisons for establishing statistical significance is based on the average correlations over cross validated test sets (\overline{cc}_{test}) is due to issues regarding the curse of dimensionality. We are considering cases where potentially $n \ll \min(p, q)$. Even with shrinkage induced by the penalization scheme, very high correlations are likely to be found when there is no relation between the two data sets. By forcing the canonical vectors to act on data they were not trained on, we avoid the overfitting common with small n , large p situations. Only if the canonical vectors are picking up on actual signal are we then likely to see a similarly high canonical correlation on the test data. Without this modification, we have found that it is nearly impossible to distinguish signal from noise.

3 Simulations

3.1 Simulation Set-up

In order to assess resistant multiple sparse canonical correlation (RMSCCA), we set up simulations with and without heavy tails (representing realistic noisy data). For each simulation of sample size n , we generate one dataset (\mathbf{x}) using a multivariate normal. Then a second dataset (\mathbf{y}) is generated as a multivariate normal distribution around a linear combination of \mathbf{x} . Similar to Chalise and Fridley (2011), we let $\mathbf{X} \sim MVN_p(0, \Sigma_{XX})$. Then for each individual l , $\mathbf{Y}_l \sim MVN_q(\mu_l, \Sigma_{YY})$, where $\mu_l = \mathbf{X}_l \times B$.

The matrix \mathbf{B} determines the relationship between \mathbf{X} and \mathbf{Y} and is all zeros except in coordinates to prescribe a particular relationship. For our purposes, \mathbf{B} is given by the equation (6). Note that $1_{n \times m}$ is an $n \times m$ matrix of 1s. Similarly, $0_{n \times m}$ is an $n \times m$ matrix of 0s. The \mathbf{B} matrix allows for multivariate linear relationships between \mathbf{X} and \mathbf{Y} . The population setup gives five sets of canonical pairs. The first canonical pair is given by the relationship between first 10 dimensions of the random variable \mathbf{X} and the first 20 dimensions of the random variable \mathbf{Y} ; the second canonical pair is represented by the next 5 dimensions of \mathbf{X} and the next 5 dimensions of \mathbf{Y} variables; and so on.

$$\mathbf{B}_{p \times q} = \begin{pmatrix} 1_{10 \times 20} & 0_{10 \times 5} & 0_{10 \times 10} & 0_{10 \times 50} & 0_{10 \times 15} & 0_{10 \times q-100} \\ 0_{5 \times 20} & 1_{5 \times 5} & 0_{5 \times 10} & 0_{5 \times 50} & 0_{5 \times 15} & 0_{5 \times q-100} \\ 0_{20 \times 20} & 0_{20 \times 5} & 1_{20 \times 10} & 0_{20 \times 50} & 0_{20 \times 15} & 0_{20 \times q-100} \\ 0_{50 \times 20} & 0_{50 \times 5} & 0_{50 \times 10} & 1_{50 \times 50} & 0_{50 \times 15} & 0_{50 \times q-100} \\ 0_{15 \times 20} & 0_{15 \times 5} & 0_{15 \times 10} & 0_{15 \times 50} & 1_{15 \times 15} & 0_{15 \times q-100} \\ 0_{p-100 \times 20} & 0_{p-100 \times 5} & 0_{p-100 \times 10} & 0_{p-100 \times 50} & 0_{p-100 \times 15} & 0_{p-100 \times q-100} \end{pmatrix} \quad (6)$$

The population covariance matrices describing each of the \mathbf{X} and \mathbf{Y} random variables ($\Sigma_{\mathbf{X}\mathbf{X}}$ and $\Sigma_{\mathbf{Y}\mathbf{Y}}$) are created to establish relationships between the known canonical groups with sufficient noise (and spurious correlations) when compared to the remaining dimensions. The underlying correlation structure for each dataset is an identity matrix except at the corresponding non-zero entries of \mathbf{B} for which there is a correlation of 0.2. That is, the first 10 dimensions of the \mathbf{X} random variable have a pairwise correlations of 0.2; the next 5 dimensions of the \mathbf{X} random variable have pairwise correlations of 0.2, etc.

Because each of the correlations between the dimensions of the \mathbf{Y} random variable is given by a combination of $\Sigma_{\mathbf{Y}\mathbf{Y}}$ and the constructed relationship between \mathbf{X} and \mathbf{Y} , the variance of each \mathbf{Y} random variable needs to be moderated to create \mathbf{Y} random variables with specified correlations. We set the variance of \mathbf{Y} assuming that \mathbf{Y} is a sum of \mathbf{X} values as well as an error term. See the appendix for the derivation of the $\Sigma_{\mathbf{Y}\mathbf{Y}}$ matrix.

Clean data were simulated as above according to a multivariate normal distribution. Data with heavier tails is given using the multivariate normal set up

as above with the additional modification that each multivariate normal observation is divided by the square root of χ_2^2 random variable divided by its degrees of freedom. Specifically, each of the n p -dimensional \mathbf{x} vectors is divided by a χ_2^2 random variable divided by 2 and then used to generate the corresponding \mathbf{y} vector. We refer to the heavy tailed data as t -like data (as the original normal random deviate is neither centered at zero nor scaled to have variance one).

3.1.1 Complete Groups

Note that the structure of \mathbf{B} leads to the idea of a *complete group*. The notion of a complete group will be important to the assessment of the methods described in the paper. We define a complete group to be the set of dimensions of \mathbf{X} and of \mathbf{Y} which are correlated. In the simulation above, there are five complete groups given by \mathbf{B} in equation 6. For example, the first complete group is represented by the dimensions $\{(1, 2, \dots, 10)\}$ in \mathbf{X} & $\{(1, 2, \dots, 20)\}$ in \mathbf{Y} . An incomplete group might be, for example, $\{(1, 4, 7)\}$ in \mathbf{X} & $\{(9, 15, 20)\}$ in \mathbf{Y} . The variables would be all true positives, but the overall complete relationship would not show up as having established a complete group of parameters. The parameters of the model are the non-zero coefficients on the complete group elements only.

3.1.2 Determining Significance

As outlined in Algorithm 3, we use a permutation scheme to determine the cutoff values for a vector of canonical correlations (keeping in mind that they *decrease* across canonical pairs). We provide a graphical representation of the algorithm to determine significance of a canonical pair. Figure 1 plots the permuted correlations and observed correlations as a function of canonical pair for one simulated

dataset. The black dots are the observed canonical correlations, and the triangles represent the 0.9 quantile of the permuted correlations for a given canonical pair. For the simulated dataset shown, there are three significant canonical pairs, as at the fourth canonical pair, the observed correlation (black dot) falls below the 0.9 quantile (0.9 chosen arbitrarily to be the value of the Q cutoff) of the permuted canonical correlations (triangle).

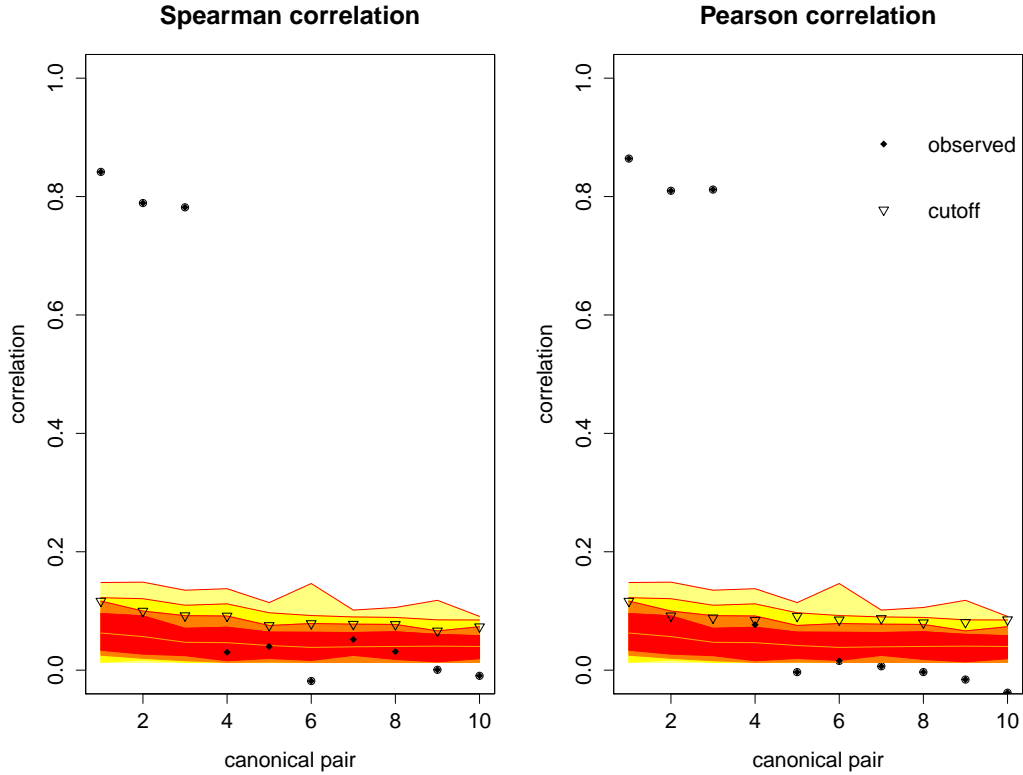


Figure 1: Given one simulated dataset, both the permuted correlations (red gradient) and observed correlations (black dot) are plotted. Additionally, the cutoff for significance is given by the 0.9 quantile of the permuted correlations (triangles) and can be seen to determine three canonical correlations (black dots) as significant. The red gradient shows additional quantiles of the permuted distribution.

3.1.3 Evaluation Metrics

In order to evaluate the methods described above, we compare the non-zero (i.e., non-sparse) coefficients to the original matrix used to generate the linear dependency between the random variables \mathbf{X} and \mathbf{Y} . Recall that the response variable is generated such that for each individual l , $\mathbf{Y}_l \sim MVN_q(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{Y}})$, where $\boldsymbol{\mu}_l = \mathbf{X}_l \times \mathbf{B}$. The matrix \mathbf{B} is given in equation (6). We detail the evaluation metrics below, keeping in mind that each of the measurements is done for those canonical pairs whose canonical correlation is above the permutation cutoff value described in Section 2.5. In the evaluation metrics below, we use the word *true* to indicate a variable which has a non-zero entry in the \mathbf{B} matrix used to simulate the data, see equation (6).

NC Pair The Number of Canonical Pairs which are significant according to the permutation test.

TPR True Positive Rate measures the total number of non-zero coefficients that are true (with double counting) divided by the sum of (total number of non-zero coefficients that are true) + (total number of empirical non-zero coefficients that are not in any complete group). That is, the ratio of total number of coefficients that are true and non-zero divided by the total number of coefficients that have empirical non-zero coefficients. The result is to measure the proportion of non-zero coefficients which are true.

TP of CG True Positive of Complete Groups gives another measure of true positives. True Positives of the Complete Groups represents the number of canonical pairs containing a complete group divided by the number of canonical pairs (NC Pair).

FN Rate The False Negative Rate measures the number of true variables with zero coefficients across all of the significant canonical pairs (out of a total number of true variables given in the model, e.g., see equation (6)).

3.2 Simulation Results

For the simulation study, we set $p = 500, q = 1000$ and let n vary along (50,100,500,1000). Each simulation was run 100 times; additionally, both λ_u and λ_v were set to range along the vector (0,0.1,0.2,0.3,0.4,0.5). By incorporating the adjusted covariance matrix into Algorithm 1, we are able to find the sparse loadings associated with each canonical pair. Additionally, each canonical pair is assessed to determine whether or not it contains a complete group. The values of the evaluation metrics above are presented below for both the clean and the t -like data across different values of the sample size.

3.2.1 Number of Canonical Pairs

The number of canonical pairs considered to be significant was determined using the permutation method (with 100 permutations) in Algorithm 3. As mentioned above, the model was set-up to have 5 canonical pairs as given in equation (6). For t -like data, both MSCCA and RMSCCA tend to give more canonical pairs than the model specifies, once the sample size is sufficiently large. This is due to the complex nature of the relationships, whereas early canonical pairs may correctly find signal in the data, they may not include every variable in the relationship (a so called complete group). Thus, later canonical pairs may be again correctly find remaining signal from the same relationship, resulting in

more estimated pairs than true pairs, though all estimates are identifying signal in the data.

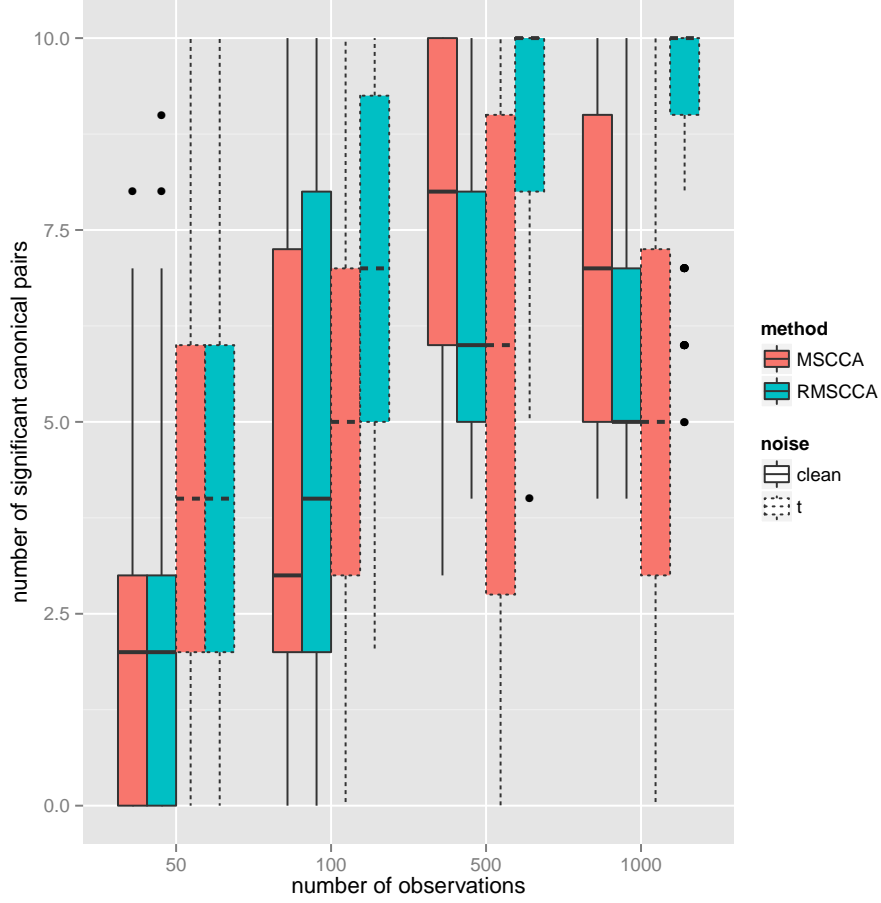


Figure 2: For each of 100 simulations, the number of canonical pairs which were determined to be significant under the permutation structure is given as a function of the simulation data size.

3.2.2 True Positive Rate

We measure true positives using two different metrics. The true positive rate (TPR) (see Figure 3) gives the proportion of non-zero coefficients across all canonical pairs. The true positive rate of complete groups (see Figure 4) gives the proportion of complete groups out of the number of canonical pairs. For

the TPR, we see a somewhat surprising result in that RMSCCA is lower across all sample sizes for t-like data. This needs to be understood in terms of Figure 4. Whereas there are indeed a higher proportion of non-zero coefficients associated with MSCAA, this is a consequence of having overly sparse solutions. For samples sizes of $n = 100$ and higher, RMSCCA has a median complete group proportion of 1, drastically outperforming its nonresistant counterpart (whose median values, across sample sizes, never exceed 0.42).

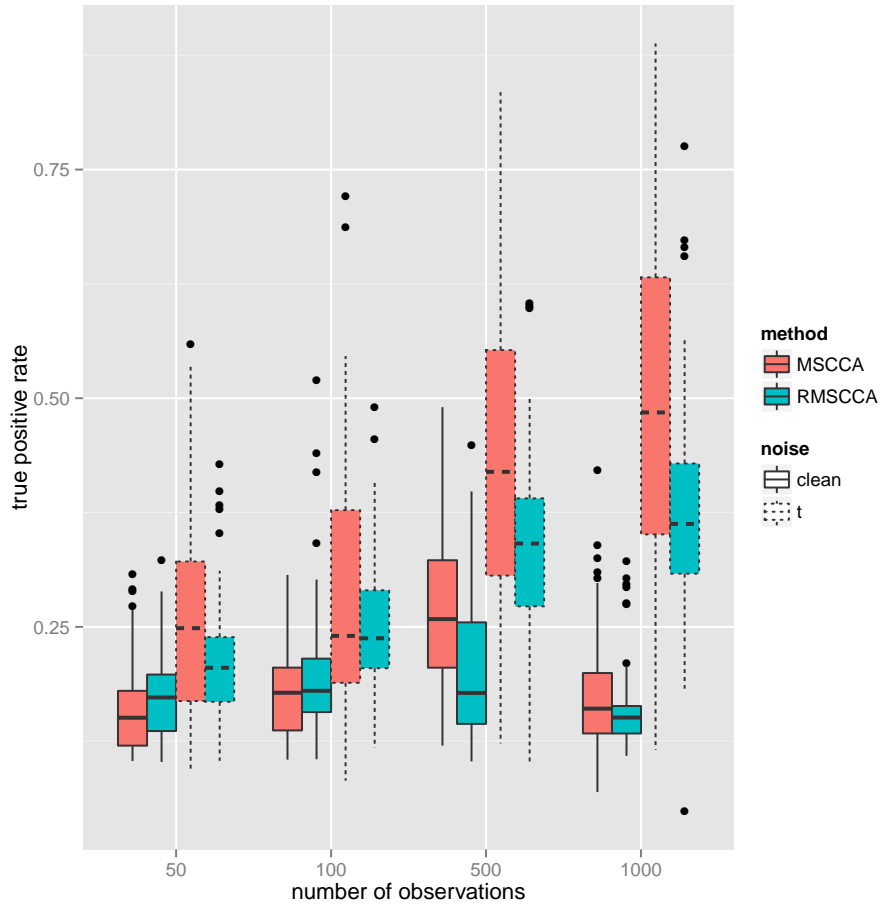


Figure 3: For each of 100 simulations, the proportion of non-zero coefficients which are true as a function of the simulation data size.

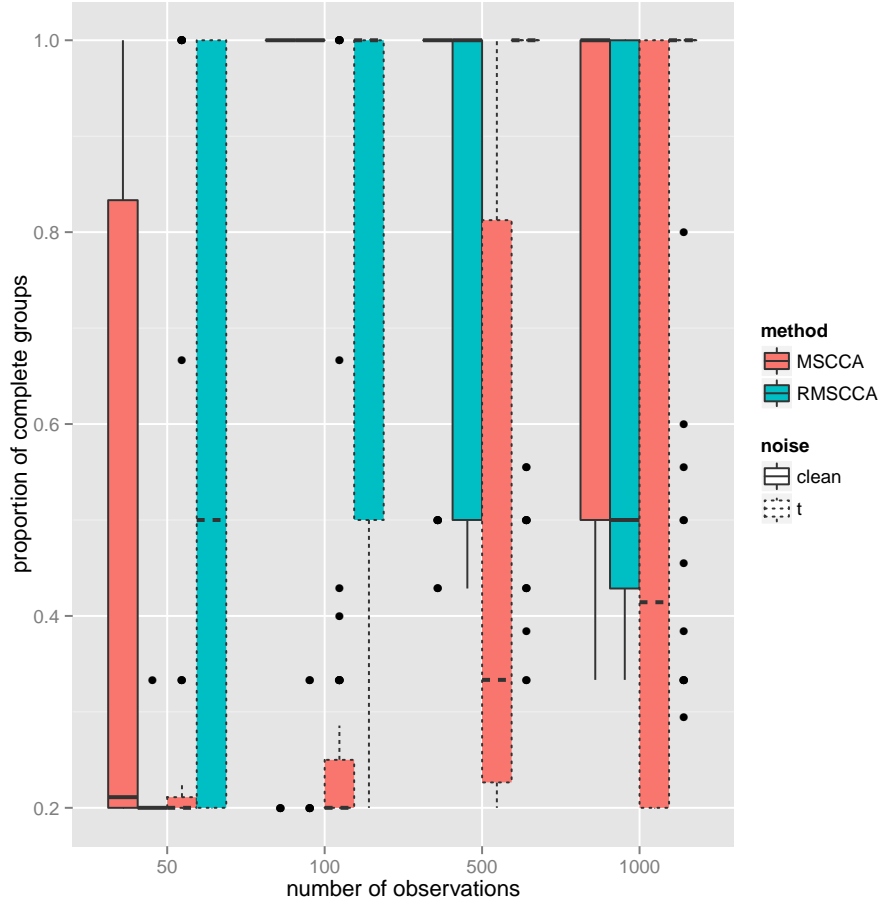


Figure 4: For each of 100 simulations, the proportion of complete groups out of the number of canonical pairs.

3.2.3 False Negative Rate

The False Negative Rate measures the proportion of true coefficients (see equation (6)) which had zero coefficients for all of the significant canonical pairs (see Figure 5). With large sample sizes, we see that only MSCCA on the t -like data has a substantial loss of power in determining positive coefficients across the significant canonical pairs. Even for lower sample sizes, RMSCCA outperforms MSCCA. (N.b., the red bar for $n = 100$ with MSCCA on t -like data is missing only due

to the small number of replications (100) in the simulation. If the box plots had been made at the 0.77 quantile instead of the 0.75 quantile, the red bar would not have disappeared.)

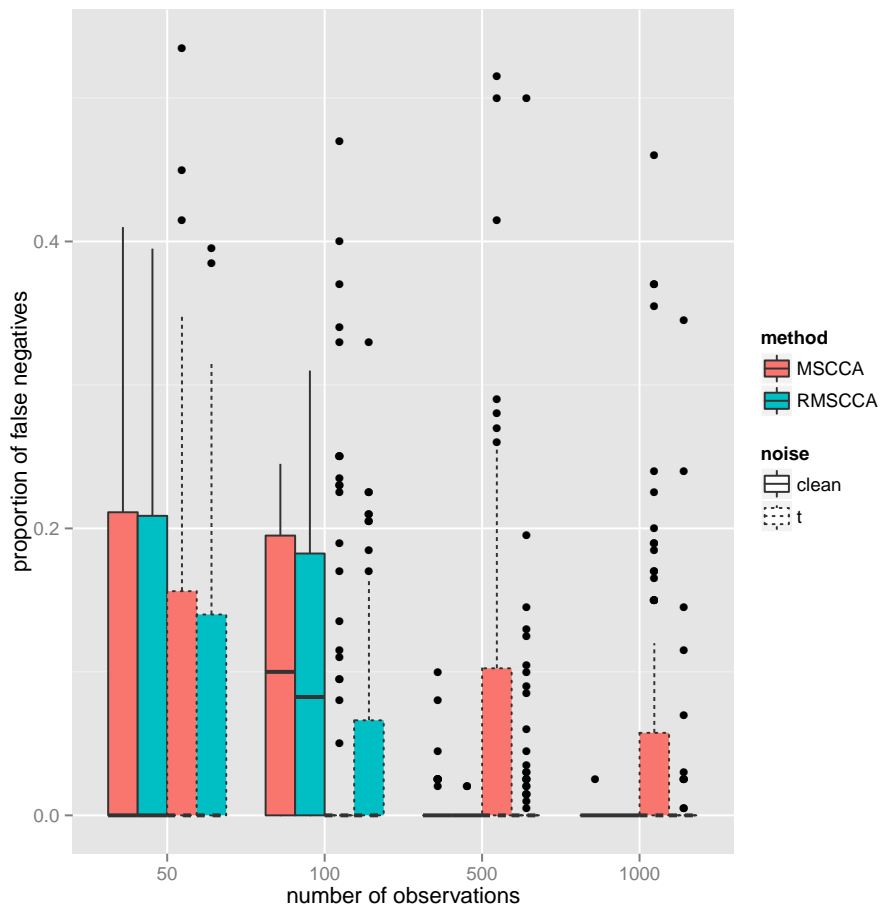


Figure 5: For each of 100 simulations, the false negative rate as a function of the simulation data structure. The clean data represent the first two boxes in each set, and the t -like data represent the second two boxes in each set.

3.2.4 Type I errors

To confirm that the familywise error rate on null data is controlled at the 0.1 level (chosen by $Q=0.9$), we simulate data with no structure between X and

Y (i.e, $B \equiv 0$). We run the complete MSCCA and RMSCCA algorithms. As above, we set $p = 500$ and $q = 1000$ letting n vary on $(50, 100, 500, 1000)$. For each of 100 simulations, we count the number of times the observed correlations are considered significant according to the permutation scheme. With null data, we expect to see the observed data above the cutoff 10% of the time because we use a 0.9 quantile cutoff. Table 1 gives the empirical type I error rates.

	Sample Size, n			
	50	100	500	1000
MSCCA	0.06	0.05	0.10	0.11
RMSCCA	0.04	0.10	0.09	0.13

Table 1: Type I error rates for data simulated as in section 3.1 with $B = 0$ so as to remove the relationship between \mathbf{X} and \mathbf{Y} . Our method accurately controls the type I error rate at 0.1.

3.2.5 Power

Power was calculated as the percent of simulation where at least one canonical correlation was above the permutation threshold. Power was calculated on all of the simulations where there was signal in the data, and so the method should have given canonical pairs above the permutation threshold. The power calculation below does not address the number of canonical pairs above the threshold.

	Sample Size, n							
	clean data				t-like data			
	50	100	500	1000	50	100	500	1000
MSCCA	0.68	0.81	1	1	0.84	0.91	0.91	0.92
RMSCCA	0.70	0.82	1	1	0.87	1.00	1.00	1.00

Table 2: Power calculations for data simulated with B so as to construct the relationships between \mathbf{X} and \mathbf{Y} described in section 3.1. The power is seen to be higher for the resistant method across the board.

3.3 Real Data

Next, we applied RMSCCA to a real biological dataset to compare our method to that of Witten et al. (2009).

We analyzed the Chin et al. (2006) copy number abnormality (CNA) and mRNA expression data available in the PMA package in R (Witten et al., 2013) in order to facilitate these comparisons.

Chin et al. (2006) measured mRNA expression of $p=19672$ genes on Affymetrix U133A microarrays and measured $q=2149$ CNAs on Bacterial Artificial Chromosome (BAC) Comparative Genomic Hybridization (CGH) arrays in aggressively treated early-stage breast tumors obtained from $n=89$ subjects. Although log-transformed CNA and microarray data are often *assumed* to follow a normal distribution, both data types are more accurately described by a heavy tailed distribution (Hardin and Wilson, 2009; Roy and Reif, 2013), and failure to account for this distribution can lead to spurious associations. Notably, many types of biological data, like genotype data, methylation data, and clinical outcomes, deviate from the assumption of normality. While non-normal data should not be accommodated by a classical SCCA algorithm, our employment of resistant estimation makes our method suitable.

Here, we applied RMSCCA to the copy number and mRNA expression data from each of the 23 chromosomes separately. Unlike Witten et al. (2009) who set their tuning parameters to achieve a sparse solution including only ≈ 25 coefficients, we used cross validation as outlined in Algorithm 2 to set our tuning parameters. Most importantly, our RMSCCA algorithm allowed us to consider multiple canonical pairs per each chromosome and to assess the significance of

these multiple canonical pairs using our permutation test as outlined in Algorithm 3 (with the maximum number of canonical pairs set to 10). The algorithm took between a few hours up to 45 hours to run for a given chromosome. The analysis was performed on a computer with two eight core AMD Opteron 6276 processors running at 1.4 GHz. The analysis can also be parallelized for a reduction in computational time.

We tested the significance for each of the top ten canonical pairs using all 23 chromosomes (compared individually). We see that the majority of the chromosomes have 10 significant canonical pairs, but not all of them. Indeed, some of the chromosomes have no significant canonical pairs, see Table 3. Though different from the analysis of Witten et al. (2009) for the reasons given above, our analysis is consistent with theirs in the sense that much of the signal within chromosomes is significant.

A closer look at chromosome 2 (using RMSCCA) shows that except for the first two canonical pairs, the test data is not significantly different from the permuted data, see figure 6. It is important to point out the purple training data points such that they suffer from the curse of dimensionality. Comparing the permuted data correlations to the training data correlations would not have been an accurate comparison due to the huge dimensionality and over-fitting that happens through the canonical correlation estimation process.

The canonical coefficients for the first canonical pairs (using RMSCCA) across each of 23 chromosomes is given in Figure 7. The red ticks represent the mRNA coefficients (both chromosomal location and magnitude of coefficient) and the green ticks represent the CNA coefficients (both chromosomal location and mag-

Chromosome	# Signif (RMSCCA)	# Signif (MSCCA)
1	10	10
2	2	4
3	3	10
4	10	7
5	10	10
6	10	10
7	10	10
8	10	10
9	10	10
10	10	1
11	10	10
12	10	10
13	10	10
14	10	10
15	10	10
16	10	10
17	0	8
18	10	10
19	10	4
20	0	2
21	10	10
22	10	10
23	0	0

Table 3: Using the breast cancer data of Chin et al. (2006), for each chromosome, the number of significant canonical pairs.

nitude of coefficient). As in the analysis by Witten et al. (2009), we see that the strong correlations (i.e., high canonical correlations) are given by comparing mRNA and CNA variables which are located at the same points along the given chromosome.

4 Conclusion

Canonical correlation analysis gives linear relationships between variables from two distinct datasets. We have extended previous work on sparse canonical cor-

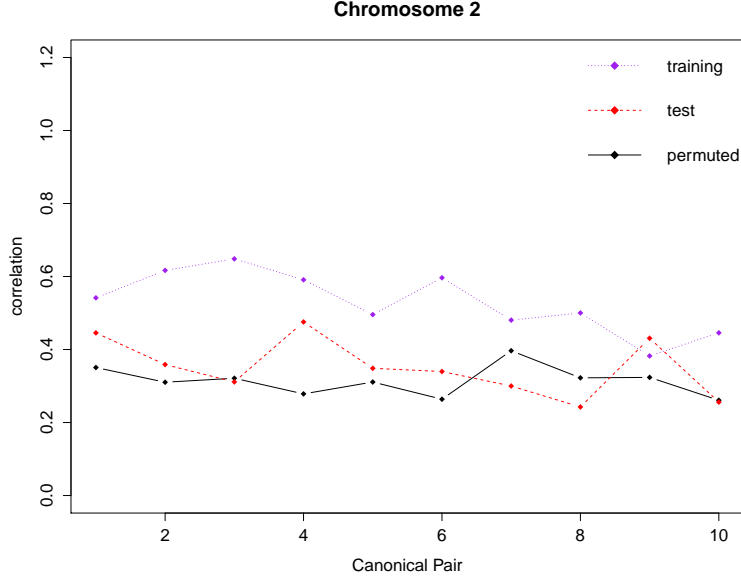


Figure 6: Using the breast cancer dataset we calculate the canonical correlation from chromosome 2 for the first 10 canonical pairs in three ways: 1. purple: canonical correlations for the full dataset based on coefficients calculated using the full dataset; 2. red: canonical correlations averaged over the cross validation procedure, coefficients from the training data, correlations on the test data; 3. the 0.9 quantile of the canonical correlations from the cross validation procedure over the *permuted* dataset.

relation to allow for both multiple canonical pairs and for resistant analysis in the setting where $n \ll p, q$. Our work shows that the method is able to find the simulated structure both in terms of number of canonical pairs and in terms of complete groups. Like MSCCA, RMSCCA still gives a large number of coefficients, but the true variables are typically returned. We see also that with heavy tailed data, it is better to use resistant correlation to avoid any leverage points. The analysis of a real dataset gives consistent results to that of Witten et al. (2009) in terms of both significance as well as connecting the chromosomal locations of the mRNA and CNA measurements.

Appendix

Consider the case of the first dimension of \mathbf{Y} , Y_1 , which is centered at the first p_1 dimensions of the random variable \mathbf{X} . Because the majority of the correlation between the dimensions of the random variable \mathbf{Y} values comes from their dependence on the random variable \mathbf{X} , let Σ_{YY} be a diagonal matrix. In contrast, Σ_{XX} is made up of $\rho(= 0.2)$ at the appropriate off diagonal elements and 1 on the diagonal.

Below is the derivation for the first diagonal entry of Σ_{YY} , $\sigma_{YY,11}$. The goal is to find $\sigma_{YY,11}$ such that $\text{cor}(y_{l1}, y_{l2}) = \rho$.

$$\mathbf{Y}_l \sim MVN_q(\mu_l, \Sigma_{YY}), \text{ where } \mu_l = \mathbf{X}_l \times B, \quad l = 1, \dots, n$$

$$\mathbf{Y}_l = \mathbf{X}_l \times B + \varepsilon_l, \text{ where } \varepsilon_l \sim MVN_q(0, \Sigma_{YY}), \quad l = 1, \dots, n$$

$$Y_{l1} = \sum_{i=1}^{p_1} X_{li} + \varepsilon_{l1}, \text{ where } \varepsilon_{l1} \stackrel{iid}{\sim} N(0, \sigma_{YY,11})$$

$$\begin{aligned} \text{Var}(Y_{l1}) &= \text{Var}\left(\sum_{i=1}^{p_1} X_{li} + \varepsilon_{l1}\right) \\ &= p_1 \sigma_{XX,11} + (p_1^2 - p_1) \sigma_{XX,12} + \text{Var}(\varepsilon_{l1}) \quad \text{WLOG} \end{aligned}$$

$$\text{Var}(Y_{l1}) = p_1 + (p_1^2 - p_1) \rho + \sigma_{YY,11}$$

$$\begin{aligned} \text{Cov}(Y_{l1}, Y_{l2}) &= \text{Cov}\left(\sum_{i=1}^{p_1} X_{li} + \varepsilon_{l1}, \sum_{i=1}^{p_1} X_{li} + \varepsilon_{l2}\right) \\ &= p_1 \sigma_{XX,11} + p_1(p_1 - 1) \sigma_{XX,12} + \text{cov}(\varepsilon_{l1}, \varepsilon_{l2}) \quad \text{WLOG} \\ &= p_1 + p_1(p_1 - 1) \rho \end{aligned}$$

$$\begin{aligned} \text{Cor}(Y_{l1}, Y_{l2}) &= \frac{p_1 + (p_1^2 - p_1)\rho}{p_1 + (p_1^2 - p_1)\rho + \sigma_{Y_{l1}, l2}} = \rho \\ \sigma_{Y_{l1}, l2} &= \left(\frac{1}{\rho} - 1\right)(p_1 + (p_1^2 - p_1)\rho) \end{aligned}$$

By increasing the variance for each of the simulated \mathbf{Y} variables involved in the true linear relationships, we create correlations of ρ ($=0.2$ in our simulations) between the \mathbf{Y} variables in a group. The cross-covariance matrix between \mathbf{X} and \mathbf{Y} ($\Sigma_{\mathbf{XY}}$) is not pre-specified, but rather it is given by the relationship between $\Sigma_{\mathbf{XX}}$, $\Sigma_{\mathbf{YY}}$, and \mathbf{B} .

References

- J. Branco, C. Croux, P. Filzmoser, and R. Oliveira. Robust canonical correlations: A comparative study. *Computational Statistics*, 20:203–231, 2005.
- P. Chalise and B. Fridley. Comparison of penalty functions for sparse canonical correlation analysis. *Computational Statistics and*, 56:245–254, 2011.
- K. Chin, S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W.-L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, T. Ryder, F. Chen, H. Feiler, T. Tokuyasu, C. Kingsley, S. Dairkee, Z. Meng, K. Chew, D. Pinkel, A. Jain, B. M. Ljung, L. Esserman, D. G. Albertson, F. M. Waldman, and J. W. Gray. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541, 2006. doi: <http://dx.doi.org/10.1016/j.ccr.2006.10.009>.
- C. Dehon, P. Filzmoser, and C. Croux. *Data Analysis, Classification, and Related Methods*, chapter Robust Methods for Canonical Correlation Analysis, pages 321–326. Springer, 2000.
- C. Gao, Z. Ma, Z. Ren, and H. H. Zhou. Minimax estimation in sparse canonical correlation analysis. Submitted to the *Annals of Statistics*, 2014.
- J. Hardin and J. Wilson. A note on oligonucleotide expression values not being normally distributed. *Biostatistics*, 10:446–50, 2009.
- J. Hardin, A. Mitani, L. Hicks, and B. VanKoten. A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics*, 8:220, 2007.

- S. Hong, X. Chen, L. Jin, and M. Xiong. Canonical correlation analysis for rna-seq co-expression networks. *Nucleic Acids Research*, 41:e95, 2013. doi: 10.1093/nar/gkt145.
- H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.
- P. Huber. Projection pursuit. *The Annals of Statistics*, 13:435–525, 1985.
- G. Karmel. Robust canonical correlation and correspondence analysis. *The Frontiers of Statistical Scientific and Industrial Applications*, 2:335–354, 1991.
- K.-A. Lê Cao, P. G. Martin, C. Robert-Granié, and P. Bess. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10:34, 2009. doi: 10.1186/1471-2105-10-34.
- D. Nguyen and D. M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18:39–50, 2001.
- E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical*, 8:1–36, 2009.
- K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 11:559–572, 1901.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL <http://www.R-project.org/>.
- P. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.
- S. Roy and A. M. Reif. Evaluation of calling algorithms for array-cgh. *Frontiers in Genetics*, 4:217, 2013.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B.*, 58:267–288, 1996.
- Y. R. Wang, K. Jiang, L. J. Feldman, P. J. Bickel, and H. Huang. Inferring gene association networks using sparse canonical correlation analysis. Submitted to the Annals of Applied Statistics, 2014.
- D. Witten and R. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications to Genetics and Molecular Biology*, 8:901–929, 2009.

- D. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.
- D. Witten, R. Tibshirani, S. Gross, and B. Narasimhan. *PMA: Penalized Multivariate Analysis*, 2013. URL <http://CRAN.R-project.org/package=PMA>. R package version 1.0.9.
- H. Wold. *Multivariate Analysis II*, chapter Nonlinear Iterative Partial Least Squares (NIPALS) Modeling: Some Current Developments, pages 383–407. New York: Academic Press, 1973.
- H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:262–286, 2006.

Chin et al. 2006 CNA/mRNA RSCCA

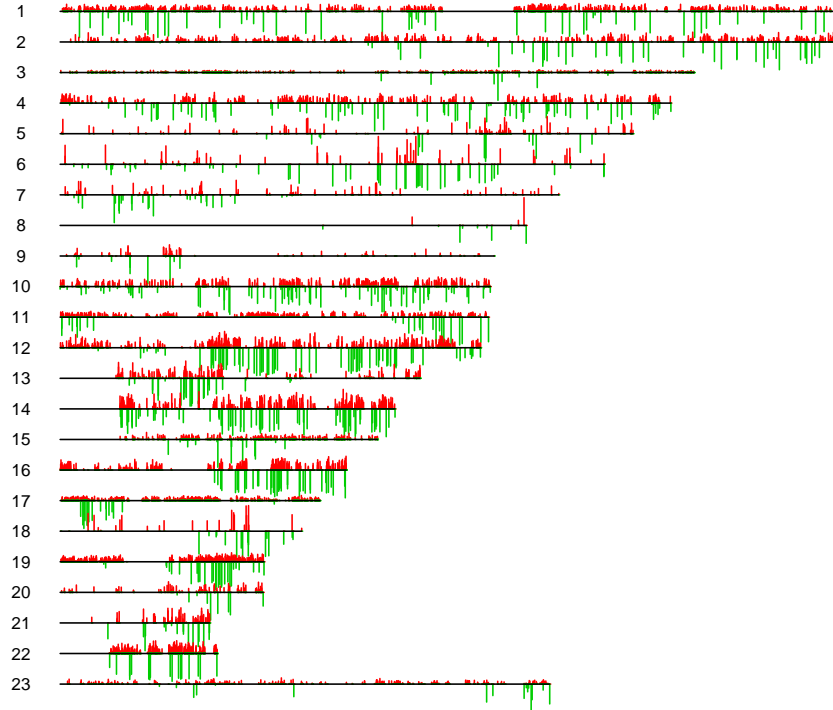


Figure 7: Using the breast cancer dataset of Chin et al. (2006), we provide the canonical coefficients for the first canonical pair across each of the 23 chromosomes. The red ticks (ticks above the horizontal line) represent the canonical coefficients associated with the mRNA data and the green ticks (ticks below the horizontal line) represent the canonical coefficients associated with the CNA data. For each tick, its location is given by the placement along the chromosome and the length of the tick is proportional to the magnitude of the canonical coefficient value.